Moreno-Marcos, Pedro Manuel; Alario-Hoyos, Carlos; Muñoz-Merino, Pedro J.; Estévez-Ayres, Iria; Delgado Kloos, Carlos; (2018). Sentiment analysis in MOOCs: A case study. *Proceedings of 2018 IEEE Global Engineering Education Conference (EDUCON2018), 17-20 April 2018, Santa Cruz de Tenerife, Canary Islands, Spain*, pp.: 1489-1496.

DOI: https://doi.org/10.1109/EDUCON.2018.8363409

Sentiment Analysis in MOOCs: A case study

Pedro Manuel Moreno-Marcos, Carlos Alario-Hoyos, Pedro J. Muñoz-Merino,
Iria Estévez-Ayres, and Carlos Delgado Kloos
Department of Telematic Engineering
Universidad Carlos III de Madrid (UC3M)
Leganés (Madrid), Spain
{pemoreno, calario, pedmume, ayres, cdk}@it.uc3m.es

***Abstract*— Forum messages in MOOCs (Massive Open Online Courses) are the most important source of information about the social interactions happening in these courses. Forum messages can be analyzed to detect patterns and learners' behaviors. Particularly, sentiment analysis (e.g., classification in positive and negative messages) can be used as a first step for identifying complex emotions, such as excitement, frustration or boredom. The aim of this work is to compare different machine learning algorithms for sentiment analysis, using a real case study to check how the results can provide information about learners' emotions or patterns in the MOOC. Both supervised and unsupervised (lexicon-based) algorithms were used for the sentiment analysis. The best approaches found were Random Forest and one lexicon based method, which used dictionaries of words. The analysis of the case study also showed an evolution of the positivity over time with the best moment at the beginning of the course and the worst near the deadlines of peer-review assessments.**

Keywords—sentiment analysis; MOOCs; learners' behavior; learning analytics; machine learning

I. INTRODUCTION

In MOOCs (Massive Open Online Courses), learners can establish social interactions with other users to ask or answer questions, with no or little support of teachers. Although there are different ways of communication, including social networks, like Facebook or Twitter, the bulk of the social interactions usually occur in the course forum [1]. In a MOOC, the number of enrolled users can be huge (it is possible to find courses with up to 100,000 learners) and, although most of them enroll just to explore [2], the total number of forum messages can be very large, and it is not feasible to read and analyze all of them individually.

Because of that, it is interesting to automate the analysis process in order to get meaningful information from the forum messages. One possible goal could be to identify which concepts present more difficulties so that teachers can adapt the materials accordingly. Another possible goal could be to cluster students in different groups (e.g., according to their personality [3]). Furthermore, it is possible to use text mining techniques to classify the sentiments that learners show in their messages. The information related to sentiment analysis can be useful to know the affective states of learners during their learning process, which is very important, as the learning process is affected depending on their sentiments [4].

A first step towards sentiment analysis in MOOCs can be to identify if forum messages are positive or negative. This can give an insight into how learners feel with the course to be able to perform modifications aimed at increasing learners' engagement and satisfaction, which is very important to ensure the success of the MOOC [5]. However, this is only the first step because more complex emotions can be detected from those sentiments, such as excitement, frustration or boredom. One possible related interest is to analyze if there are specific moments in the course where overall learners' positivity decreases, which could be related to some materials or activities causing troubles to learners.

Sentiment analysis has been carried out in other fields, such as movie reviews or tweets [6], but there are very few contributions in the area of MOOCs, with exceptions such as the work by Wen, Yang and Rosé, who analyzed the relationship between a ratio based on the positive and negative terms in the posts and dropout [7]. Besides, there is not a clear approach of how to tackle sentiment analysis in MOOCs nor a comparison of different techniques.

This work aims to address this issue and provide a comprehensive comparison between different machine learning approaches for the particular case of detecting the polarity (i.e. if a message is positive or negative) in the forum messages. Furthermore, as the final intention is to be able to discover patterns based on the analysis, a real MOOC is used as a case study to detect behaviors that learners show based on their sentiments.

The structure of the paper is as follows. Section 2 provides a background on what has been researched on prediction in MOOCs, and particularly with forum data, and other approaches related with sentiment analysis. Section 3 describes the methodology used to carry out this work, including the description of the dataset, the variables, techniques and metrics used. Section 4 presents the results and discussion of the study, and finally, the main conclusions and future research directions are indicated in Section 5.

II. RELATED WORK

The sentiment analysis is used to detect different affective states. These states or patterns in general can be detected

through different techniques, such as process mining (e.g. [8]) or discourse analysis. The discourse analysis can make use of prediction techniques, which are used in this work in order to infer affective states.

Apart from predicting affective states in MOOCs, similar techniques have been applied on predicting dropouts (because of its high rates), forecasting if the learner is going to pass the course or not (or similarly if the learner is going to receive a certificate), or the score the learner is going to achieve. In such cases, different variables have been used, mainly related to the platform use (e.g. inactivity times [9]), forum activity, video-watching activities [10], and the results of previous assignments. For example, Ren, Rangwala and Johri [11] predicted intermediate assignment grades and found that the number of previous quizzes attempted had the strongest correlation with the score. Similarly, Sinha and Cassel [12] predicted grades, classifying them into different categories: low, medium, high and very high achievement.

In most of the articles already mentioned, one common component is that forum variables are usually considered in the analysis, to a greater or lesser extent. Furthermore, as forums are the main source for social interactions in MOOCs, there have also been contributions in the analysis of the social component in MOOCs. Some examples are the identification of the top contributors in the forum [13], the personality of learners [3], users' confusions [13] or if there will be intervention from an instructor in a forum message or not [15]. Apart from that, there have been contributions with the aim of classifying messages. For example, Brinton et al. [16] analyzed 73 MOOCs from Coursera to classify messages according to their relevance using algorithms based on HITS (Hyperlink-Induced Topic Search) and TF-IDF (Term Frequency – Inverse Document Frequency), and observed the decline of forum participation over time.

In the same line of classifying messages in MOOCs, only a few contributions related to sentiment analysis can be found. For example, Ramesh, Kumar, Foulds, and Getoor [17] proposed a classification of posts (using Markov random fields) according to different course aspects, which included videos, quizzes, social interactions or the certificate, and a fine classification of each aspect (e.g. video messages were classified in those related to the video quality, audio, subtitles, etc.). They also classified messages depending on their sentiments, which could be positive negative or neutral. Besides, Bakharia [18] presented a preliminary work which covered sentiment analysis of learners based on unigram (n-gram of size 1) features.

Apart from that, there are also contributions which included variables regarding learners' sentiments to measure the relationship with dropout, but without providing an evaluation of how the sentiment analysis algorithm works (using defined metrics), unlike in the aforementioned examples. For example, Chaplot, Rhim and Kim [19] used a lexicon-approach in which the sentiment of forum posts was the sum of the sentiment scores of all the words that were provided by *SentiWordNet 3.0* [20]. Similarly, Tucker, Pursel and Divinsky [21] assigned words to positive and negative emotions (each one in the range 1-5) and found that students' sentiments were slightly and positively correlated with quiz performance, while strongly and negatively correlated with homework assignments.

As it has been shown, there are very few contributions that focus specifically on sentiment analysis on MOOCs (although there could be examples that use simples approaches for other purposes). Because of that, it is also interesting to review what has been done in other contexts. Piryani, Madhavi and Singh [22] conducted a review on Opinion Mining and Sentiment Analysis research and some conclusions were that machine learning (supervised) approaches dominated (67.20%) over the lexicon (unsupervised) approaches, which are the most common in the contributions in MOOCs. Besides, the most frequent datasets were about reviews (e.g. movie reviews), followed by tweets and news articles.

As reference articles, both supervised and unsupervised examples are presented. As a supervised approach, Pang and Lee [23] compared three machine learning (supervised) algorithms: Naïve Bayes, maximum entropy classification and Support Vector Machines (SVM). These algorithms were chosen because their philosophies were different and they had been used in previous text categorization studies. In that study, they found SVM as the best algorithm but acknowledged the fact that the sentiment classification problem is more challenging than other classification problems because sometimes the message could be written in an ironic tone, and there is certain subjectivity in the classification process. Similarly, Boiri and Moens [24] used the same algorithms using unigrams as binary features (occurrence or nonoccurrence of the feature).

As an unsupervised approach, Turney [25] also classified reviews using the difference of the mutual information between the phrase and the words "excellent" and "poor", with reported accuracies between 0.66 (movie reviews) and 0.84 (automobile reviews) depending on the context, which is also relevant for this problem. Besides, Hu and Liu [26] presented an early contribution where they computed the sentences' polarity by extracting the opinion words (adjectives) and looking for them in dictionaries of positive and negative words, taking into account the possible effect of negation words. Similarly, Li and Wu [27] included a list of modifiers denoting the emotional intensities to enhance the model. A different approach seen on the literature [28] consists on building a tree with the words of the sentences and applying different defined rules to get the polarity. This method provided competitive performance although the complexity was higher than in previous techniques.

Therefore, there are many ways to perform sentiment analysis but there are very few contributions in MOOCs which perform an evaluation of the approaches. This work will include a novel contribution in the field, providing a comparison of different techniques, including new adaptations of lexicon-based unsupervised algorithms. Besides, previous articles mainly focused on analyzing the relationship with dropouts, but this work will provide a different analysis of results, with the aim to discover and discuss patterns in learners' behaviors that may be useful for a later enhancement of the course.

III. METHODOLOGY

A. Dataset

This study was carried out using the data of the first edition of the MOOC called *Introduction to Programming with Java – Part 1: Starting to Program in Java*. This was the first one of a trilogy of MOOCs developed by Universidad Carlos III de Madrid (UC3M) for learning Java from scratch. The course was hosted on edX and launched between April and June 2015. The MOOC, taught in English, was structured in five weeks where learners had to complete weekly assignments, which were either graded tests or programming tasks (peer assessments). In total, there were 95,555 enrolled users, although only 5,126 of them contributed at least once in the forum. This subset of learners produced 13,302 messages, which are the posts that will be used in the study.

The data used for carrying out the research was provided directly from edX. Particularly, one file contained the forum messages and the characteristics of the course discussion interactions (e.g. votes, number of replies, timestamps, etc.). This file is named with format *{org}-{course}-{run}-{site}.mongo,* and was retrieved from the Database Data [29].

B. Approach

The sentiment analysis conducted in this study is focused on determining the polarity of learners' sentiments in messages, that is, whether these are positive or negative messages (initially neutral messages are excluded to consider the problem as a binary classification one). As it has been discussed in Section 2, it is possible to use a lexicon (unsupervised) or machine learning (supervised) approach. In this work, the intention is to compare both types of techniques.

One limitation of machine learning techniques is that they need data for training. As edX only provides raw data, messages had to be labeled manually. This was a very time-consuming task, and 500 posts have been labelled, which is a limited number but representative enough for the case study, as even smaller datasets have been used in the literature (100 reviews were considered in [30]). Nonetheless, these labelled messages were needed to evaluate the models regardless they are supervised or not. Therefore, the labelling task was needed although only lexicon methods were used.

Apart from that, it is worth mentioning that the labelling process is very subjective and there were messages that could be classified with the opposite label (positive or negative) by other people. As messages were only labelled by one person, messages were also flagged to indicate the degree of confidence of the person labelling them (if the message is flagged, it means that the polarity is clear). This allows providing two possible values when reporting the results: one value when all messages are used (R1) and another when only flagged messages are used (R2). The benefit of having two values is that it is possible to have an interval which takes into account the subjectivity of the process.

Regarding the evaluation of results, Leave-One-Out cross validation was used with the 500 labelled messages. As this set is limited, Leave One-Out was preferred to have as many samples as possible in the training set. As for the metrics, Pélanek [31] stated that for predicting affective states, such as polarity, boredom, concentration, confusion and so on, the kappa coefficient is mainly used and accepted. Because of that, it will be used to measure the degree of agreement between the person who labelled the messages and the algorithm, avoiding the effect produced by chance. Besides, Pelánek stated that the Area Under the Curve (AUC) can be also appropriate in this context, unlike the accuracy. For this reason, results will be also reported with AUC. Regarding the interpretation of those metrics, guidelines followed in the literature ([32] for AUC and [33] for kappa) will be used.

IV. RESULTS AND DISCUSSION

This section is divided in three main parts. The first two cover the different lexicon (unsupervised) and machine learning (supervised) approaches, while the third one presents the conclusions obtained in the case study.

A. Lexicon approaches (unsupervised)

Lexicon approaches for sentiment analysis try to use the information of the polarity that a language has. The advantage of these methods is that it is possible to perform sentiment analysis without training data. However, specific vocabularies for each language are required. The first proposal makes use of dictionaries of positive and negative English words [26] available for free[1] and computes a variable related to the orientation (or polarity) of the message depending on if words are positive or negative. Algorithm 1 shows the pseudocode of the algorithm.

There is a general loop to iterate for every word of the message (after the tokenization process for separating the words of the post). Emoticons are also considered, since they usually express a polarity; if an emoticon is found (from those listed in Table I), the polarity is what it represents (e.g. a message with a smiley-face is considered positive). If not, it is checked if the word is a negation word (see Table I). In positive case, a flag is set and a window of five words (this number was also used by Hu and Liu [26]) is enabled in which the polarity of words is reversed. For example, if the positive word "good" is found in this window, it will be considered as negative). It is important to mention the trade-off between performance and computation time. There can be more complex approaches to identify the scope of negation [34], but their complexity means more time to compute the polarity. In this case, it has been preferred to avoid large computation times to make the process of getting the polarity feasible for the massive number of messages in the MOOC.

After checking emoticons and negation words, if none of previous cases have happened, words are looked up in dictionaries of positive and negative words. If a word is found in a positive lexicon, the orientation variable will increase (or decrease if the negation flag is active) one unit, and vice versa for words found in a negative lexicon. To be able to normalize

[1] Source: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

the results afterwards, as not all words appear in dictionaries, there is a variable which stores the number of words which contributed to the orientation. Then, if the negation window is enabled, the window value will be decreased and disabled when reaching 0. When all words have been processed, the results are normalized and adjusted to be in range between 0 (negative) and 1 (positive) to facilitate the computation of the AUC with different thresholds.

ALGORITHM 1: Lexicon approach for sentiment analysis

```
words ← tokenize(message)
orientation ← 0
negation_flag ← false
negation_window ← 5
tokens_total ← 0
for each word in words, do
    if word is a positive emoticon, then
        orientation ← 1
        tokens_total ← tokens_total + 1
        break
    end if
    if word is a negative emoticon, then
        orientation ← 0
        tokens_total ← tokens_total + 1
        break
    end if
    if word is a negation word, then
        tokens_total ← tokens_total + 1
        negation_flag ← true
        negation_window ← 5
        orientation ← orientation – 1
    end if
    if word is a positive word, then
        tokens_total ← tokens_total + 1
        if negation_flag is true
            orientation ← orientation – 1
        else
            orientation ← orientation + 1
        end if
    end if
    if word is a negative word, then
        tokens_total ← tokens_total + 1
        if negation_flag is true
            orientation ← orientation + 1
        else
            orientation ← orientation – 1
        end if
    end if
    negative_window ← negative_window – 1
    if negative_window is 0, then
        negation_flag ← false
    end if
end
normalize orientation
adjust orientation in range 0-1
```

TABLE I. EMOTICONS AND NEGATION WORDS

Type	Emoticon / Word
Positive emoticons [35]	`:D`, `:)`, `:-)`, `:-D` and `;D`
Negative emoticons [35]	`:(`, `:'(`, `:-(`, `;(`, `>:(` and `=(`
Negation words [36]	Not, no, yet, never, nowhere, nobody, none, nothing, hardly and scarcely

TABLE II. RESULTS OF CLASSIFICATION WITH LEXICON APPROACHES

Method	AUC		Kappa	
	R1	***R2***	***R1***	***R2***
Dictionaries	**0.71**	**0.78**	**0.38**	**0.54**
SentiWordNet	0.65	0.75	0.24	0.52

The second approach to predict the polarity of messages was similar, but used *SentiWordNet 3.0.0* instead of the dictionaries of words. *SentiWordNet* is a lexical resource, publicly available[2], designed for sentiment analysis and opinion mining applications. It is the result of automatically annotating all *WordNet synsets* (sets of synonyms of English words) according to their degrees of positivity, negativity and neutrality [20]. For each of the words, two values (range 0-1) are shown, which correspond to the positive and negative polarity. In this case, words are also separated by their lexical category and it is necessary to indicate it when looking up words. The categories available are: adjective (a), noun (n), adverb (r) and verb (v).

The algorithm used in this case is similar to the one presented previously, but with two main differences. The first one is that instead of adding/subtracting one unit, the *SentiWordNet* score of the word will be used. The second difference is that it is necessary to obtain the lexical category to look up words, although this task can be easily done with libraries, such as *OpenNLP*[3].

Table II summarizes the results where two different values are presented. The first one (R1) is obtained after using all labelled messages (n=500) and the second one (R2) is obtained using only the posts that had a clearer polarity (n=134). Results show that although the second algorithm (*SentiWordNet)* is more complex, it provides worse results. The first algorithm (Dictionaries) provides a fair AUC and moderate kappa, which are reasonable taking into account that in sentiment analysis, results are always lower than in other classification tasks because of the subjectivity of the process.

[2] Source: http://sentiwordnet.isti.cnr.it

[3] Source: https://opennlp.apache.org/docs/1.8.0/manual/opennlp.html

This subjectivity is also visible when comparing results of both values (R1 and R2), where differences of values imply that almost neutral posts (those not included in R1 because they may not be clear) are difficult to classify.

B. Supervised approaches

In the previous subsection, two different unsupervised approaches were presented. However, it is possible to take advantage of the power of machine learning algorithms to try to improve the classification results. In this subsection five different techniques will be compared and will be also contrasted with the unsupervised approaches.

First of all, to apply machine learning algorithms, it is necessary to define some indicators that will be used for training the models. In this case, the proposal is based on forum-related variables obtained from the file provided by edX. These variables were obtained taking into account the most important fields that edX provides regarding forum interactions (e.g. votes, type of message), considering previous contributions (e.g. unigrams) and taking advantage of what had been implemented in the dictionaries approach (orientation variable). The list of those variables is as follows:

- *Number of positive votes*. It is a continuous variable which indicates the number of positive votes the message received.
- *Endorsed message*. It is a categorical variable which indicates if the message has been flagged by the instructor or the message originator because of its value and relevance.
- *Message length*. It is a continuous variable which indicates the size (in characters) of the message.
- *Orientation*. It is a continuous variable (with range 0-1) whose value depends on the positive and negative words found in the message, using the algorithm presented in Algorithm 1.
- *Type of message*. It is a categorical variable which indicates if the message is the first of the thread (known as *CommentThread*) or if it a first-level or second-level response (known as *Comment*).
- *Number of responses*. It is a continuous variable which indicates the number of replies a message had.
- *Day of the course*. It is a continuous variable which indicates the number of days between the day when the message was posted and the beginning of the course.
- *Unigrams*. They are different categorical variables which indicate if different words appear or not in the message. These words are: *problem*, *assessment*, *points*, *date*, *thanks*, *great*, *agree*, *luck* and *!* (exclamation mark). These words were selected because of their possible relationship with positive or negative messages. For example, a message with the word *great* is likely to be positive.

TABLE III. RESULTS OF CLASSIFICATION WITH SUPERVISED (S) AND UNSUPERVISED (U) ALGORITHMS

Method	AUC		Kappa	
	R1	***R2***	***R1***	***R2***
Logistic Regression (S)	0.68	0.84	0.18	**0.61**
SVM (S)	0.70	0.77	0.08	0.42
Decision Trees (S)	0.64	0.74	0.27	0.48
Random Forest (S)	**0.71**	0.82	0.28	0.47
Naïve Bayes (S)	0.66	**0.85**	0.21	0.58
Dictionaries (U)	**0.71**	0.78	**0.38**	0.54
SentiWordNet (U)	0.65	0.75	0.24	0.52

These variables have been computed at a message level and then they have been fed to different algorithms [37][38]:

- *Logistic Regression*.
- *Support Vector Machines (SVM)*.
- *Decision Trees*.
- *Random Forest*.
- *Naïve Bayes*

The selection of algorithms was done taking into account what algorithms have been used more frequently in contributions related to prediction in MOOCs. All these algorithms have been tested using the Python library *scikit-learn*[4], which includes a comprehensive implementation of different algorithms and tools for data mining and data analysis. The results obtained with these techniques are presented in Table III. The best results are remarked in bold font.

Results show that among the supervised approaches, Random Forest provides the best results both for AUC and kappa when considering all the labeled messages (R1), and achieves also acceptable results with posts that have a clear polarity (R2), with a good AUC and a moderate kappa. In contrast, Naïve Bayes obtains very good results when the polarity of the messages is clear (R2), with AUC of 0.85 and kappa of 0.58, but it is not recommendable in all situations, since it offers worse results when using all messages (R1). Similarly, logistic regression can obtain the best kappa (0.61) in R2 but fails when using all the messages (R1). In the case of SVM and decision trees, their performance is always worse than Random Forest, which can be considered the most reliable option of the supervised approaches. However, when comparing Random Forest with the dictionaries approach, the performance of both algorithms is very similar, which means that both ways to handle the sentiment analysis of forum messages are recommendable in this MOOC.

As for the values obtained, it is not possible to compare them directly with other articles in the state of the art since the contexts are different, the used datasets are different, and, in most cases, they do not refer to the area of education or MOOCs; still it is interesting to check what others have achieved to take them as a reference. In MOOCs, the only

[4] Source: http://scikit-learn.org/stable/documentation.html

article which evaluated results classifying messages in positive or negative achieved accuracies over 0.79 [18]. However, in that article, the proportion of positive messages was always above 75% and in one case it was 93%, which means the accuracy could reach 0.93 by classifying all posts with the predominating class (positive). Because of that, the accuracy is not the best metric, although it is widely used, and AUC and kappa have been preferred in this work.

Out of the area of MOOCs, Fernández-Gavilanes, Álvarez-López, Juncal-Martínez, Costa-Montenegro, and González-Castaño [28] compared different contributions which used sentiment analysis for movie reviews or tweets related to a political debate, and found accuracies between 0.54 and 0.77. In this work, despite accuracies have not been used because of the aforementioned problems, they have been computed to be able to compare. Results show that the best accuracy obtained is 0.82 for R2 with logistic regression, and 0.74 for R1 with the dictionaries approach. This means that the results presented here are reasonably good (the best accuracy is better than the 0.77 achieved in the abovementioned work), although it is important to point out that the type of data and the context can cause variations in the results, and that results from other contributions can only be considered as a reference.

C. *Analysis of the case study*

Different approaches for sentiment analysis were compared in previous sections. However, the final goal of these techniques is to be able to detect patterns on how learners behave in the course. Because of that, this section will provide some visualizations obtained from the sentiment analysis, and their corresponding discussion. The dictionaries approach will be used for the visualizations as it provides the best results when using all messages. The neutral category will be included for those messages whose orientation is 0.5 (in the range 0-1), so that the most positive and negative messages are identified, while the ones that are less clear are just reported as neutral.

In the Java MOOC, there were 13,302 posts in total. Among them, 5,292 were classified as positive 2,934 as negative and 5,076 as neutral. This means that if including only positive and negative message, 64.33% of posts were classified as positive. This entails that, in general, the forum is constructive and there is a positive feeling among learners, though there is a significant number of negative messages, which are probably critical to different aspects of the course. As shown in Figure 1, which depicts the distribution of positive, negative and neutral messages over time, there are more messages on the initial days, followed by a decrease in the number of messages per day, which is consistent with other contributions in the literature [39]. It is also interesting to see that there are some peaks were the amount of positive/negative messages increase or decrease. These peaks can be better appreciated in Figure 2 where there is a curve with the percentage of positive messages (considering only positive and negative messages, and excluding neutral ones to analyze the relationship between only positivity/negativity).

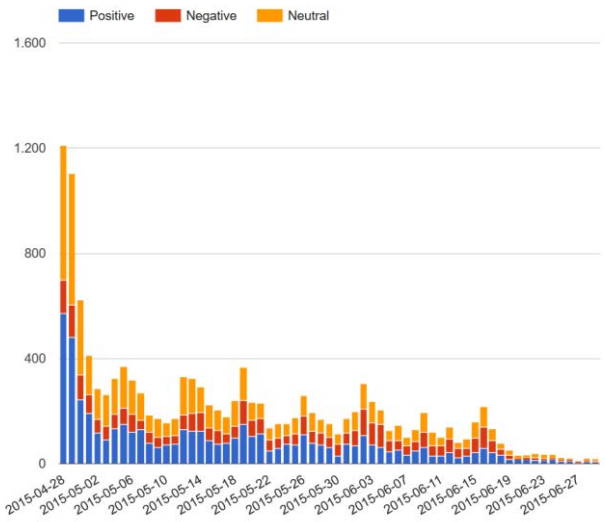

Fig. 2. Histogram with the distribution of positive, negative and neutral messages over time in the MOOC forum.

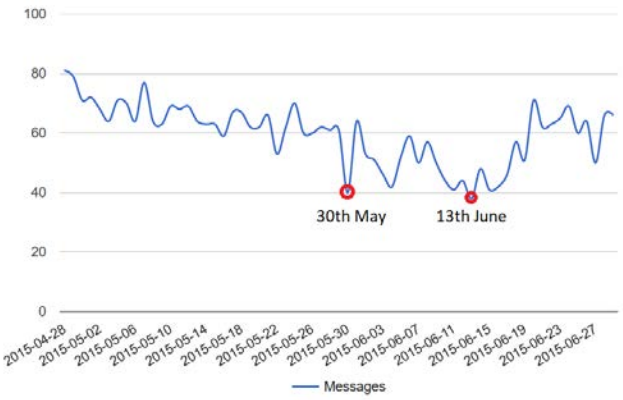

Fig. 1. Percentage of positive messages in the forum over time (including positive and negative messages, excluding neutral messages).

This figure shows that the most positive trends are at the beginning of the course. This is reasonable as people tend to be more enthusiastic about the course in the first stages [40], as it happens in this course. The figure also shows that the percentage of positivity is decreasing with certain peaks towards the middle of the course, with a prominent negative peak on May 30. This date corresponds to three days before the deadline of the third graded test and the first programming task (which was graded through peer assessment). When looking for messages in that date, it was found that there were many negative messages by learners who reported that their code did not work, which helps explain the results.

The same effect was reported on June 13, which is also three days before the deadline of the second programming task (also peer assessed). Therefore, it seems that learners tend to be more negative when they face programming tasks that are taken into account to calculate the final grade. However, it is

an interesting finding that the worst moments are three days before the submission deadline. Considering that deadlines were on Tuesdays, this means that learners tended to do the activities on the weekend as it is maybe easier to find time there for the MOOC. Another observation is that after the last deadline (June 16), the positivity raises significantly. The possible explanation is that learners are happier because the course is finished, and they managed to reach the end, particularly those who obtained a passing grade.

V. CONCLUSIONS

The analysis has shown the results obtained by using different algorithms to carry out sentiment analysis with the messages taken from a MOOC forum. Initially, two lexicon (unsupervised) approaches were described, which were based on the use of dictionaries of words and *SentiWordNet*. Next, five machine learning algorithms (supervised) were used. Results showed that the most reliable supervised approach was Random Forest, while the dictionaries method had also good behavior. This is consistent with other contributions (although in other contexts), where Random Forest was the best algorithm or was among the best ones [41][42]. The best value for AUC was between 0.71 and 0.85 and the best value for kappa was between 0.38 and 0.61, depending on the data sample used for evaluation. These results are reasonable taking into account that in sentiment analysis it is not possible to achieve as high rates as in other classification tasks, such as dropout prediction, because of the subjectivity of the process. Besides, results were also contrasted with other contributions in the literature and taking them as a reference, it can also be said that the results obtained in this work are acceptable.

After the overall analysis with both unsupervised and supervised algorithms, the dictionary of words was selected to get further insights on learners' sentiments in the forums throughout the course, and particularly in a real MOOC about Java programming. Results showed that learners were more positive at the beginning of the course and that their positivity decreased over time. This suggests that learners are initally motivated but as the course evolve, they find difficulties than can influence their polarity. Besides, some negative peaks were found near the deadline of the programming assignments, which were related to the reporting problems in their code in those dates. In addition, it was observed that the negative peaks were produced on Saturdays, which means that although deadlines were on Tuesdays, learners tended to use the weekends for doing the MOOC activities.

Despite being able to obtain the abovementioned conclusions, this work is not exempt of some limitations that are worth mentioning. The most important one was that there were no labelled data for training the algorithms and for their evaluation, and this required a manual labelling process. Besides, only one person did the labelling and, as the polarity can be subjective in some occasions, there might be discrepancies if other people had labelled the same messages. Furthermore, as this task was very time consuming, there was a limited subset of 500 labelled messages, which is enough for the purpose of this work but still limited.

Another important limitation was that all messages used for training and evaluation were taken from the same MOOC. This entails that the results are valid in this context and probably valid in similar courses, but conclusions may have some variations when transferring to other courses. Because of that, one possible future work would be collecting data from other MOOCs to train with a wider variety of messages. This would make it possible to use the models in different contexts (avoiding overfitting) and to analyze if it is possible to enhance the predictive power.

Similarly, it would be interesting to analyze MOOCs in other areas of knowledge to identify if learners' behaviors are similar or not. A possible example would be a MOOC related to humanities where the tasks and target users can be completely different. Besides, as a conclusion was that sentiments were more negative near the deadlines, it would be useful to analyze a self-paced course where there are not intermediate deadlines to see if there is any noticeable pattern. Finally, the most important aspect would be to analyze how all the techniques presented and discussed here, their results and the patterns identified based on the data can be used in a pedagogical way to improve the learning processes and to support learners to improve knowledge acquisition. In this line, a possible application of this work could be the identification of the parts of the course where sentiments are worse to support instructors to reflect on the possible issues. Moreover, results could be applied for further work on detecting affective states, which can be useful to design proper strategies to increase student's engagement and retention in the course.

ACKNOWLEDGMENT

This work has been co-funded by the Madrid Regional Government, through the eMadrid Excellence Network (S2013/ICE-2715), by the European Commission through Erasmus+ projects MOOC-Maker (561533-EPP-1-2015-1-ESEPPKA2-CBHE-JP), SHEILA (562080-EPP-1-2015-1-BEEPPKA3-PI-FORWARD), and LALA (586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), and by the Spanish Ministry of Economy and Competitiveness, projects SNOLA (TIN2015-71669-REDT), RESET (TIN2014-53199-C3-1-R) and Smartlet (TIN2017-85179-C3-1-R). The latter is financed by the State Research Agency in Spain (AEI) and the European Regional Development Fund (FEDER). It has also been supported by the Spanish Ministry of Education, Culture and Sport, under a FPU fellowship (FPU016/00526).

REFERENCES

[1] S. Mak, R. Williams, and J. Mackness, "Blogs and forums as communication and learning tools in a MOOC," Proc. International Conference on Networked Learning (NLC '10), pp. 275-285, May 2010.

[2] J. Whitehill, J.J. Williams, G. Lopez, C.A. Coleman, and J. Reich, "Beyond prediction: First steps toward automatic intervention in MOOC student stopout," SSRN Electronic Journal, May 2015.

[3] G. Chen, D. Davis, C. Hauff, and G.J. Houben, "On the impact of personality in massive open online learning," Proc. Conference on User

Modeling Adaption and Personalization (UMAP '16), pp. 121-130, Jul. 2016.

[4] B. Grawemeyer, M. Mavrikis, W. Holmes, A. Hansen, K. Loibl, and S. Gutiérrez-Santos, "The impact of feedback on students' affective states," Proc. International Workshop on Affect, Meta-Affect, Data and Learning (AMADL '15), vol. 7, pp. 4-13, Jun. 2015.

[5] T. Phan, S.G. McNeil, and B.R. Robin, "Students' patterns of engagement and course performance in a Massive Open Online Course," Computers & Education, vol. 95, pp. 36-44, Apr. 2016.

[6] C.N. Dos Santos and M. Gatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts," Proc. International Conference on Computational Linguistics (COLING '14), pp. 69-78, Aug. 2014.

[7] M. Wen, D. Yang, and C. Rose, "Sentiment Analysis in MOOC Discussion Forums: What does it tell us?," Proc. International Conference on Educational Data Mining (EDM '14), Jul. 2014.

[8] D. Leony, P.J. Muñoz-Merino, J.A. Ruipérez-Valiente, A. Pardo, and C. Delgado Kloos, "Detection and Evaluation of Emotions in Massive Open Online Courses," Journal of Universal Computer Science, vol. 21, no. 5, pp. 638-655, May 2015.

[9] S. Halawa, D. Greene, and J. Mitchell, "Dropout prediction in MOOCs using learner activity features," Experiences and best practices in and around MOOCs, vol. 7, pp. 3-12, Mar. 2014.

[10] C. Ye and G. Biswas, "Early Prediction of Student Dropout and Performance in MOOCSs Using Higher Granularity Temporal Information," Journal of Learning Analytics, vol. 1, no. 3, pp. 169-172, 2014.

[11] Z. Ren, H. Rangwala, and A. Johri, "Predicting performance on MOOC assessments using multi-regression models," arXiv preprint arXiv:1605.02269, 2016.

[12] T. Sinha, and J. Cassell, "Connecting the Dots: Predicting Student Grade Sequences from Bursty MOOC Interactions over Time," Proc. ACM Conference on Learning@ Scale (L@S '15), pp. 249-252, Mar. 2015.

[13] C. Alario-Hoyos, P.J. Muñoz-Merino, M. Pérez-Sanagustín, C. Delgado Kloos, and H.A. Parada G, "Who are the top contributors in a MOOC? Relating participants' performance and contributions," Journal of Computer Assisted Learning, vol. 32, no. 3, pp. 232-243, Jun. 2016.

[14] D. Yang, R. Kraut, and C.P. Rosé, "Exploring the Effect of Student Confusion in Massive Open Online Courses," Journal of Educational Data Mining, vol. 8, no. 1, pp. 52-83, 2016.

[15] S. Chaturvedi, D. Goldwasser, and H. Daumé III, "Predicting Instructor's Intervention in MOOC forums," Proc. Annual Meeting of the Association for Computational Linguistics (ACL '14), pp. 1501-1511, Jun. 2014.

[16] C.G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F.M.F. Wong, "Learning about social learning in MOOCs: From statistical analysis to generative model," IEEE transactions on Learning Technologies, vol. 7, no. 4, pp. 346-359, Oct/Dec 2014.

[17] A. Ramesh, S.H. Kumar, J.R. Foulds, and L. Getoor, "Weakly Supervised Models of Aspect-Sentiment for Online Course Discussion Forums," Proc. Annual Meeting of the Association for Computational Linguistics (ACL '15), pp. 74-83, Jul. 2015.

[18] A. Bakharia, "Towards Cross-domain MOOC Forum Post Classification," Proc. ACM Conference on Learning@ Scale (L@S '16), pp. 253-256, Apr. 2016.

[19] D.S. Chaplot, E. Rhim, and J. Kim, "Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks," Proc. International Conference on Artificial Intelligence in Education (AIED '15), Jun. 2015.

[20] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," Proc. Conference on Language Resources and Evaluation (LREC '10), vol. 10, pp. 2200-2204, May 2010.

[21] C. Tucker, B.K. Pursel, and A. Divinsky, "Mining student-generated textual data in MOOCs and quantifying their effects on student performance and learning outcomes," Computers in Education Journal, vol. 5, no. 4, pp. 84-95, Oct/Dec 2014.

[22] R. Piryani, D. Madhavi, and V.K. Singh, "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015," Information Processing & Management, vol. 53, no. 1, pp. 122-150, Jan. 2017.

[23] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," Proc. ACL Conference on Empirical Methods in Natural Language Processing (EMNLP '02), vol. 10, pp. 79-86, Jul. 2002.

[24] E. Boiy, and M.F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," Information retrieval, vol. 12, no. 5, pp. 526-558, Oct. 2009.

[25] P.D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," Proc. Annual Meeting on Association for Computational Llinguistics (ACL '02), pp. 417-424, Jul. 2002.

[26] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), pp. 168-177, Aug. 2004.

[27] N. Li and D..D Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," Decision support systems, vol. 48, no. 2, pp. 354-368, Jan. 2010.

[28] M. Fernández-Gavilanes, T. Álvarez-López, J. Juncal-Martínez, E. Costa-Montenegro, and F.J González-Castaño, "Unsupervised method for sentiment analysis in online texts," Expert Systems with Applications, vol. 58, pp. 57-75, Oct. 2016.

[29] edX, "EdX Research Guide Release," https://media.readthedocs.org/pdf/devdata/latest/devdata.pdf. 2017.

[30] J. Singh, G. Singh, and R. Singh, "Optimization of sentiment analysis using machine learning classifiers," Human-centric Computing and Information Sciences, vol. 7, no. 32, Dec. 2017.

[31] R. Pelánek, "Metrics for evaluation of student models," Journal of Educational Data Mining, vol. 7, no. 2, pp. 1-19, 2015.

[32] A.D. Mezaour, "Filtering Web Documents for a Thematic Warehouse Case Study: eDot a Food Risk Data Warehouse (extended)," Proc. International IIS Intelligent Information Processing and Web Mining (IIPWM '05), pp. 269-278, Jun. 2005.

[33] J.R. Landis, and G.G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no.1, pp. 159-174, Mar. 1977.

[34] A. Asmi, and T. Ishaya, "Negation identification and calculation in sentiment analysis," Proc. International Conference on Advances in Information Mining and Management (IMMM '12), pp. 1-7, Oct. 2012.

[35] H. Wang, and J.A. Castanon, "Sentiment expression via emoticons on social media," Proc. IEEE International Conference on Big Data (Big Data '15), pp. 2404-2408, Oct/Nov 2015.

[36] H. Johnson, "English Sentence Negation: How to Negate Sentences in English", http://www.linguisticsgirl.com/english-sentence-negation-how-to-negate-sentences-in-english/. 2013.

[37] M. Hammad, and M. Al-awadi, "Sentiment Analysis for Arabic Reviews in Social Networks Using Machine Learning," Proc. International Conference on Information Technology: New Generations (ITNG ' 16), pp. 131-139, Apr. 2016.

[38] N.F.F. da Silva, E.R. Hruschka, and E.R. Hruschka Jr., "Tweet sentiment analysis with classifier ensembles," Decision Support Systems, vol. 66, pp. 170-179, Oct. 2014.

[39] C. Alario-Hoyos, M. Pérez-Sanagustín, C. Delgado-Kloos, and M. Munoz-Organero, "Delving into participants' profiles and use of social tools in MOOCs," IEEE Transactions on Learning Technologies, vol. 7, no. 3, pp. 260-266, Jul/Sep 2014.

[40] H. Nilsen, "Influence on Student Academic Behaviour through Motivation, Self-Efficacy and Value-Expectation: An Action Research Project to Improve Learning," Issues in Informing Science & Information Technology, vol. 6, pp. 545-556, May 2009.

[41] X. Fang, and J. Zhan, "Sentiment analysis using product review data," Journal of Big Data, vol. 2, no. 5, Dec. 2015.

[42] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera, "Opinion mining and sentiment analysis on a twitter data stream," Proc. International Conference on Advances in ICT for emerging regions (ICTer '12), pp. 182-188, Dec. 2012.